Synopsis on

A STUDY ON APPLICATIONS OF INFORMATION THEORETIC MEASURES IN DATA MINING

AUGUST-2020

Submitted for registration in the degree of

Doctor of Philosophy

DEPARTMENT OF MATHEMATICS

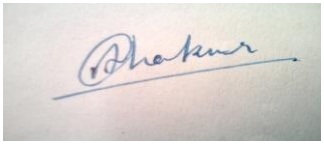
CHITKARA UNIVERSITY

HIMACHAL PRADESH

Submitted by

KOMAL KUMARI DESWAL (PHDASM18057)

Under the joint supervision of



Dr. Ashok Kumar
(Chitkara University)

Dr.Krishan Kumar
(DCRUST, Sonipat)

INDEX

Sr. No.	Topic	Page Number
1.	Introduction	
2.	Literature Review	
3.	Importance of proposed Research Investigation(motivation)	
4.	Objectives	
5.	Methodology	
6.	Outcomes	
7.	Work Plan	
8.	References	

1. Introduction

1.1 Data Mining

Data mining is a rapidly growing field with the aim of analyzing big data in academia and industry. In it information-theoretical methods play a key role in discovering useful knowledge from a large amount of data. For example, probabilistic modeling of data sources based on information-theoretical methods such as maximum entropy, the minimum description length principle, rate-distortion theory, Kolmogorov complexity, etc. have turned to be very effective in machine learning problems in data mining such as model selection, regression, clustering, classification, structural/relational learning, association/causality analysis, transfer learning, change/anomaly detection, stream data mining, sparse modeling, etc. As real data become complex, further advanced information-theoretical methods are currently emerging to adapt data in realistic sources such as non-i.d. sources, heterogeneous sources, network type data sources, sparse sources, etc. Information-theoretical data mining methods have successfully been applied to a wide range of application areas including finance, education, marketing, intelligent transportation systems, multi-media processing, health care, network science, etc. In the recent year, huge amounts of data being collected and stored in databases everywhere across the globe mainly come from information industry and social sites. These are needed to extract and classify for useful information and knowledge from large data. The data mining is most popular knowledge acquisition technique, which deals with this problem. It is used to extract the required data from large databases and perform the automated extraction and generate the predictive information from large database. It is actually the process of finding the hidden information or patterns from the repositories Data mining consists of the various multi-disciplinary skill based technical approaches including machine learning, statistics, artificial intelligence and database system. The goal of the data mining process is to discover knowledge from large databases and transform into a human understandable format. It is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data. The insights derived via Data Mining can be used for marketing, fraud detection, and scientific discovery, etc. Data mining is also called as Knowledge discovery, Knowledge extraction, data/pattern analysis, information harvesting, etc.

1.2 Data Mining Implementation Process

Decision Tree Mining is a type of data mining technique that is used to build Classification Models. It builds classification models in the form of a tree-like structure, just like its name. This type of mining belongs to supervised class learning. In supervised learning, the target result is already known. Decision trees can be used for both categorical and numerical data. The categorical data represent gender, marital status, etc. while the numerical data represent age, temperature, etc.

Figure 1: Data Mining Process

1.3 Information Theory

Information Theory is relatively a new branch of applied mathematics which was made mathematical rigorous only in 1940. In broader sense, it deals with the study of problems concerning any system which includes information processing regression, relational learning, clustering information storage, information retrieval and decision making. More precisely, it deals with all theoretical problems of transmission of information over communication channels. Further, it includes the study of uncertainty measures, and practical as well as economical methods of coding information to be transmitted.

The first studies in this direction were undertaken by Nyquist (1924, 1928) and Hartley (1928) who introduced the entropy of a distribution of equally probable events. In Shannon (1948) published a landmark paper, which laid the foundation of the modern day's information theory. He introduced a measure of information or entropy for a general finite complete probability distribution and gave a characterization theorem of the entropy measure introduced by him.

Entropy as defined by Shannon and added upon by other physicists is closely related to thermo dynamical entropy. Entropy is randomness and how much information a message contains is measured by the extent it combats entropy. The second half of the 20^{th} century was characterized by the tremendous development of systems in which the transmitted information is coded in a digital form. By this coding, the real nature of the information signal becomes secondary, i.e. the same system can transmit simultaneously signals for very different nature like data, audio, video etc. This development has been made possible by use of more powerful integrated circuits. Although it is mainly during the last 25 years that the truly operational digital systems have been developed, the theoretical foundations for all these developments date back to the work of Shannon and others in the mid of the 20^{th} century which led to the development of information theory as a field of mathematics. Information Theory basically considers following three fundamental issues related to 'compression', 'lossy data compression' and 'channel capacity'.

The theory is concerned with the mathematical laws governing systems designed to communicate information. It sets up quantitative measures of information and of the capacity of various systems to transmit, store and process information. It overlaps with communication theory but is oriented towards the fundamental limitations on the processing and communication of information instead of the detailed operation of the devices employed. In past sixty years, the literature on information theory has grown quite voluminous and apart from its applications in communication theory it is being used in retrieval and intelligence gathering. Also it has found deep application in many fields and remains an active area of research, and in the information age, provides insight and direction for the design and analysis of various systems particularly communication and computer systems.

2. Literature Review

2.1 Branch of Mathematics

As discussed, the development in the field of Information Theory started rigorously as a branch of Mathematics in 1948 after the remarkable work of *C.E. Shannon (1948)*. Being an electrical engineer his goal was to get maximum line capacity with minimum distortion.

An American mathematician, electronic engineer and cryptographer, well known as "the father of information theory", had provided the base to the field related to information known as “Information Theory”.

2.2 Shannon Entropy

Information is the key feature of Shannon’s information theory. According to him, information occurs only if there is some prior uncertainty and is equal to the amount by which the uncertainty has been reduced from an experiment/observation.

Shannon conceived the statistical nature of the communication signal with that of the random variable $\boldsymbol{X = (X_1, X_2, X_3, \dots, X_n)}$ having probability distribution $\boldsymbol{P = (p_1, p_2, p_3, \dots, p_n)}$, where ‘$\boldsymbol{p_i}$’ is the probability at some ‘$\boldsymbol{X = X_i}$’ and introduced the *measure of information (or uncertainty*) as

$$H(X) = -\sum p_i \log p_i \qquad (1)$$

such that $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$ associated with this experiment. This is also called the ***Shannon’s measure of entropy***.

Properties of the Uncertainty Function

a) **Non-negativity:** $H(P)$ is always non-negative, that is, $H(P) = -\sum p_i \log p_i \geq 0$

b) **Maxima:** $H(p_1, p_2, \dots, p_n) \leq \log n$ with equality when $p_i = \frac{1}{n}$ for all i.

c) **Continuity:** $H(p_1, p_2, \dots, p_n)$ is continuous function of $p_i's$.

d) **Symmetry:** $H(p_1, p_2, \dots, p_n)$ is a symmetric function of $p_i's$ i.e. it is invariant with respect to the order of the outcomes.

e) **Additivity:** If $P = \{p_1, p_2, \dots, p_n\}$ and $Q = \{q_1, q_2, \dots, q_n\}$ are two independent probability distributions, then we have, $H(P * Q) = H(P) + H(Q)$, where $P * Q$ is the joint probability distribution.

In case, X is a continuous random variable, e.g. if it denotes the lifetime of a ‘unit’ with p.d.f $f(x)$ then the measure of Shannon uncertainty associated with X is given by

$$H(X) = -\int_0^\infty f(x)\, log\, f(x)\, dx \qquad (2)$$

This is a measure of uncertainty of the lifetime of a unit. The idea is that a unit with higher uncertainty is less reliable than a unit with low uncertainty.

2.3 Generalization

1) An additive generalization of the measure Shannon Entropy was given by Renyi’s (1961) as

$$H^\alpha(X) = \frac{1}{1-\alpha} \log[\sum p^\alpha];\ \alpha \neq 1, \alpha > 0 \qquad (3)$$

It has similar properties as that of Shannon Entropy, but it contains additional parameter ‘α’ which can be used to make it more or less sensitive to the shape of probability distributions [10]. If it has large positive value this measure is more sensitive to events that occur often, while for large negative values of ‘α’, it is more sensitive to the events which happen seldom. It should be worthwhile to control this tradeoff explicitly, as in many cases it may be important to distinguish weak signal overlapping with much stronger one. Entropy measures that depend on powers of probability, like Renyi Entropy, provide such control.

2) A non-additive generalization of the measure (1) is given by Havrda and Charvat (1967)

$$H_\alpha(X) = \frac{1}{2^{1-\alpha}-1} [\sum (p_i)^\alpha - 1];\ \alpha \neq 1, \alpha > 0 \qquad (4)$$

When $\alpha \to 1$, both measure (3) and (4) reduces to (1).

It satisfies the non-additivity

$H(X * Y) = H(X) + H(Y) + (2^{1-\alpha} - 1)H(X)H(Y)$

It is more general than Shannon entropy and simpler than Renyi entropy.

3) Belis and Guiasu (1968) extended the concept of Shannon entropy to the systems with *quantitative and qualitative* characteristics and gave the '*useful*' information measure

$$H(X; U) = -\sum u_i \, p_i \log p_i \quad (5)$$

Where 'u_i' is the utility, or the cost factor assigned with the output $= x_i$.

4) Recently, Rao et al. (2004) introduced an alternative measure of uncertainty called Cumulative Residual Entropy (CRE) of a random variable X, defined as

$$\xi(X) = -\sum P(X > x) \log P(X > x)$$

This measure is based on cumulative distribution function (CDF) rather than probability density, and is thus, in general more stable, since the distribution function is more regular because it is defined in an integral form unlike the density function, which is defined as the derivative of the distribution.

5) Fuzzy entropy –based relevance measure, in information theory, the Shannon entropy measure is generally used to characterize the impurities of a collection of samples .Assuming X as a discrete random variable with a finite set of n elements, where $X = \{x1, x2, x3, \dots, xn\}$,then if an element xi occurs with probability p(xi)[20],the entropy H(X) of X is defined as follows:

$H(X) = \sum_{i=1}^{n} p(xi) log_2 \, p(xi)$

Where n denotes the n denotes the number of elements .An extension of Shannon entropy with fuzzy sets, which is used to support the evaluation of entropies, is called fuzzy entropy. It was introduced in 1972 [21], after which a number of modifications were introduced to the original fuzzy entropy method [22, 23]

The proposed fuzzy entropy method is based on the utilization of the Fuzzy C-Means Clustering algorithms (FCM), which is used to construct the membership function of all features. The data may be belong to two or more clusters simultaneously and the belonging of a data point to the clusters is governed by the membership values[24].Similar data points are placed in the same cluster and dissimilar data points normally belong to different clusters. The membership values of the data points are reorganized iteratively to reduce the dissimilarity of two data points.

2.4 Applications

The Shannon's measure of entropy finds its applications in coding theory, marketing and in almost all other social and physical sciences. For example, in coding theory if $\bar{l} = \sum_{i=1}^{n} l_i p_i$ is the average code word length of a base D uniquely decipherable code for the random variable $X = (X_1, X_2, X_3, \dots, X_n)$ with probability distribution $P = (p_1, p_2, p_3, \dots, p_n)$, then

$$-\sum p_i \log p_i \leq \bar{l} \leq -\sum p_i \log p_i + 1$$

The Shannon measure (1) has been characterized and generalized by many researchers like refer [3] Cover and Thomas (1991), [1] Aczel, J. and Daroczy, Z. (1975) and [8] Renyi (1961). Among different application areas of Shannon's Entropy, an active area of current research is in data mining**.**

3. Decision Tree

The Decision tree is one of the classification techniques which is done by the splitting criteria. The decision tree is a flow chart like a tree structure that classifies instances by sorting them based on the feature (attribute) value. Each node in a

decision tree represents a feature in an instance to be classified. All branches denote an outcome of the test, each leaf node hold the class label. The instances are classified from starting based on their feature value. Decision tree generates the rule for the classification of the data set. Three basic algorithms are widely used that are ID3, C4.5, and CART [9]. ID3 is an iterative Dichotomer 3. It is an older decision tree algorithm introduced by Quinlan Ross in 1986[5] . The basic concept is to make a decision tree by using the top-down greedy approach. C4.5 is the decision tree algorithm generated by Quinlan, J.R [8]. It is an extension of ID3 algorithm. C4.5 algorithm is widely used because of its quick classification and high precision. CART stands for Classification Regression Tree introduced by Bremen [8] .The property of CART is that it is able to generate the regression tree. In Regression tree leaf node contains a real number instance of a Class. A decision tree classifier is built in two phases as Growth phase and Prune phase. After the preliminary tree has been built that is ‘growth phase', a sub-tree is created with the least estimated error rate, that is the 'prune phase'. The process of pruning the preliminary tree consists of removing small, deep nodes of the tree resulting from 'noise' contained in the training sample thus decreasing the risk of 'over fitting' and ensuring in a more precise classification of unknown data. As the decision tree is being built, the goal at each node is to decide the split attribute (feature) and the split point that best divides the training instances belonging to that leaf. The value of a split point depends on how well it separates the classes. Numerous splitting indices have been proposed in the precedent to evaluate the quality of the split.

4. Justification for Research

4.1 Motivation

Entropy Computation is used to create compact decision trees with successful classification. The size of the decision tree, the performance of the classifier is based on the entropy calculation. So the most precise entropy can be applied to the particular classification problem. The different entropies based approach can be applied in any classification problem. Such as detecting faults in industrial application, Medical diagnosis, loan approval, pattern recognition, classifying market trends etc. The work under consideration is a comparative study based on various entropies like Shannon, R´enyi, quadratic, Havrda and Charvt, Taneja entropy and it also builds a model that takes some specific entropies in parallel and produce more precise classification for data set and a result of this classification is comparable with the other machine learning techniques. This entropy based approach can be applied in real world classification problems.

DISCUSSION ON RESEARCH PAPERS

Author(s)	Algorithm used	Contribution(s)	Method
Yi-bin et al, 2017	ID3	Improvement of ID3 Algorithm Based on Simplified Information Entropy and Coordination Degree	Approximation using Maclaurin formula
Wang et al,	ID3	Novel approach to select the	Rough set

2017		splitting attribute.	
Soni and Pawar, 2017.	ID3	Novel feature selection approach for decision tree construction	Attribute importance combined with information gain
Man et al, 2018	Random Forest	Optimization of decision tree algorithm node splitting Adaptive Parameter Selection	Adaptive Parameter Selection
Fang et al, 2017	ID3	Introduction of the concept of mutual information in selecting the splitting attribute instead of information gain	Mutual information
Chen et al, 2013 Improving Iteration process	C4.5	Classification accuracy of C4.5 decision tree algorithm based on sample selection	Iteration process
Muslim et al, 2018	C4.5	Improving the accuracy of the algorithm to predict credit receipts	Split Feature reduction model and bagging ensemble
Cherfi et al, 2018	C4.5	A novel algorithm for building smaller and more accurate decision trees	Arithmetic mean and median
Muslim et al, 2017	C4.5	Optimization of C4.5 algorithm for breast cancer diagnosis	Particle Swarm Optimization

4.2 Research Gaps

The revolution in database technologies has resulted in an increase of data accumulation in many areas, such as financial, marketing and the biological and medical sciences. It has become crucial to locate hidden information by scrutinizing these data effectively. Data mining techniques have been discussed widely and applied successfully in the areas of medical research, scientific analysis and business applications. Recently, the absorption of data mining techniques in medical diagnosis has provided new insights in a large number of medical applications. Feature selection has many advantages such as shortening the number of measurements, reducing the execution time and improving transparency and compactness of the suggested diagnosis.

5. Problem Statement

Feature selection aims to reduce the amount of unnecessary, irrelevant and redundant features. It helps retrieve the most relevant features in datasets and improves the classification accuracy with less computational effort. If the features are not chosen well, even the best classifier performs poorly. The work aims to improve feature relevance measures based on Shannon entropy values and devise

best feature selection strategies. The intention is to select the correct set of features for classification when datasets contains noisy, redundant and vague information.

6. Research Objectives

I propose to study information theoretic measures and their applications specifically with the following objectives:

- To study applications of various existing information theoretic measures in the classification of Data Mining.
- To analyse the performance of various existing classification algorithms.
- To propose and study some new Technique based on information theoretic measures to boost performance of classification algorithms.
- To compare and analysis of existing methods with proposed technique to improve accuracy and reduce time.

7. Methodology

To find information from the disordered data, decision tree with classification algorithm is best approach. Decision tree is used to build classification and regression models. It is used to create data models that will predict class labels or values for the decision-making process. Using a decision tree, we can visualize the decisions that make it easy to understand and thus it is a popular data mining technique.

7.1 Working of Decision Tree

A decision tree is a supervised learning algorithm that works for both discrete and continuous variables. It splits the dataset into subsets on the basis of the most significant attribute in the dataset. How the decision tree identifies this attribute and how this splitting is done is decided by the algorithms.

The most significant predictor is designated as the root node, splitting is done to form sub-nodes called decision nodes, and the nodes which do not split further are terminal or leaf nodes.

In the decision tree, the dataset is divided into homogeneous and non-overlapping regions. It follows a top-down approach as the top region presents all the observations at a single place which splits into two or more branches that further split. This approach is also called a *greedy approach* as it only considers the current node between the worked on without focusing on the future nodes.

The decision tree algorithms will continue running until a stop criteria such as the minimum number of observations etc. is reached.

Once a decision tree is built, many nodes may represent outliers or noisy data. Tree pruning method is applied to remove unwanted data. This, in turn, improves the accuracy of the classification model.

To find the accuracy of the model, a test set consisting of test tuples and class labels is used. The percentages of the test set tuples are correctly classified by the model to identify the accuracy of the model. If the model is found to be accurate then it is used to classify the data tuples for which the class labels are not known.

Some of the decision tree algorithms include ID3, C4.5, and CART. My research work will be based upon classification algorithms like ID3, C4.5 etc. because other than high accuracy their performance can be enhanced using various information theoretic measures.

7.2 Decision Tree Induction

Decision tree induction is the method of learning the decision trees from the training set. The training set consists of attributes and class labels. Applications of decision tree induction include astronomy, financial analysis, medical diagnosis, manufacturing, and production.

A decision tree is a flowchart tree-like structure that is made from training set tuples. The dataset is broken down into smaller subsets and is present in the form of nodes of a tree. The tree structure has a root node, internal nodes or decision nodes, leaf node, and branches.

The root node is the topmost node. It represents the best attribute selected for classification. Internal nodes of the decision nodes represent a test of an attribute of the dataset leaf node or terminal node which represents the classification or decision label. The branches show the outcome of the test performed.

Some decision trees only have *binary nodes*, that mean exactly two branches of a node, while some decision trees are non-binary.

7.3 Decision Tree Induction for Machine Learning: ID3

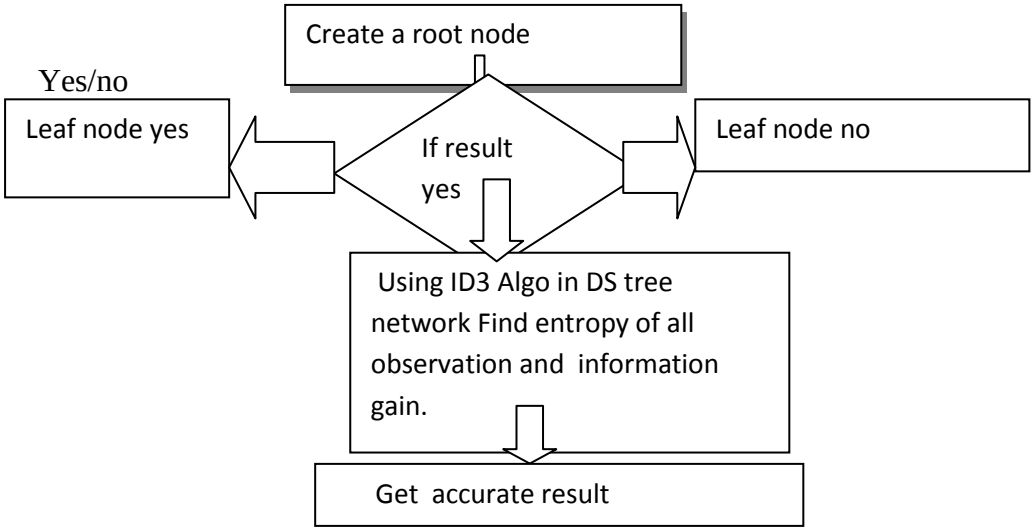
In the late 1970s and early 1980s, J.Ross Quinlan was a researcher who built a decision tree algorithm for machine learning. This algorithm is known as **ID3, Iterative Dichotomiser**. This algorithm was an extension of the concept learning systems described by E.B Hunt, J, and Marin. ID3 later came to be known as C4.5. ID3 and C4.5 follow a greedy top-down approach for constructing decision trees. The algorithm starts with a training dataset with class labels that are portioned into smaller subsets as the tree is being constructed.

Algorithm

- Initially, there are three parameters i.e. **attribute list, attribute selection method and data partition**. The attribute list describes the attributes of the training set tuples.
- The attribute selection method describes the method for selecting the best attribute for discrimination among tuples. The methods used for attribute selection can either be Information Gain or Gini Index.
- The structure of the tree (binary or non-binary) is decided by the attribute selection method.
- When constructing a decision tree, it starts as a single node representing the tuples.
- If the root node tuples represent different class labels, then it calls an attribute selection method to split or partition the tuples. The step will lead to the formation of branches and decision nodes.
- The splitting method will determine which attribute should be selected to partition the data tuples. It also determines the branches to be grown from the node according to the test outcome. The main motive of the splitting criteria is that the partition at each branch of the decision tree should represent the same class label.
- The above partitioning steps are followed recursively to form a decision tree for the training dataset tuples.
- The portioning stops only when either all the partitions are made or when the remaining tuples cannot be partitioned further.
- The complexity of the algorithm is described by **n * |D| * log |D|** where n is the number of attributes in training dataset D and |D| is the number of tuples.

Steps and data flow diagram of Decision Tree Induction method:

Step1: The first step will be to create a root node.
Step2: If all results are yes, then the leaf node "yes" will be returned else the leaf node "no" will be returned.
Step3: Find out the Entropy of all observations and entropy with attribute "x" that is E(S) and E(S, x).
Step4: Find out the information gain and select the attribute with high information gain.
Step5: Repeat the above steps until all attributes are covered.


Create a root node
Yes/no
Leaf node yes
If result
yes
Leaf node no
Using ID3 Algo in DS tree network Find entropy of all observation and information gain.
Get accurate result

Here we are using Decision Tree Classification following reasons.

Enlisted below are the various merits of Decision Tree Classification:

1. Decision tree classification does not require any domain knowledge, hence, it is appropriate for the knowledge discovery process.
2. The representation of data in the form of the tree is easily understood by humans and it is intuitive.
3. It can handle multidimensional data.
4. It is a quick process with great accuracy.

8. Expected Outcomes

Data compression and the related information-theoretic techniques, find a wide use for investigation in computational biology. Such a pervasive use has grounds in some outstanding notions that deeply characterizes data compression, in particular universality and quantification of statistical dependence via information measures. Those notions give raise to methods that need very few assumptions on the data models and, as a consequence, very minor parameter estimations for the application of those tools. That seems to be a major advantage for computational biology applications, where the statistical modeling of the data is a highly non-trivial task. In addition, the low-computational demand of those tools allows them to scale well with dataset size, even on a genomic scale. In conclusion, versatility, 'parameter-free' data and association mining, and speed are the main advantages for the use of data compression in biological investigations. However, a non-trivial organizational effort is required in order for this area to collect, in a homogenous way, the set of ideas and tools that would constitute the critical mass required to be recognized as one of the pillars in Bioinformatics. Moreover, the connection of data compression to machine learning is also receiving attention (Sculley and Brodley, 2006) and hopefully it will result in further unifying principles and methodologies, with impact on many disciplines, including the ones connected to the Life Sciences.

5. Work Plan

In the real world, we are confronted not only with complex and high dimensional data sets, but usually noisy, incomplete and uncertain data, where the application of traditional methods of knowledge discovery and of entropy for the use in knowledge discovery and data mining, with particular emphasis on open source data. We will consider some open problems and future data mining techniques on the basis of their performance and compare different types of entropy methods with a number of different purposes and possible application areas in the field of data mining.

References:

[1] Shannon,C.E. 1948, 'A mathematical theory of communication' *Bell Syst. Tech. J.*, 27, 379–423, 623–656

[2] Rényi, A. 1961, 'On measures of information and entropy' *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability* 1960. pp. 547–561.

[3] Kolmogorov A.N. 1965, 'On tables of random numbers' , *Sankhyā Ser. A*, 25, 369–375.

[4] Aczel, J. and Daroczy, Z. 1975, '*On measure of information and their characterizations*', Academic Press, New York.

[5] Quinlan, J. R. 1986, '*Induction of decision trees*'. Machine learning, 1(1), 81-106

[6] Ash, R.B. 1990, '*Information Theory*', Dover Publications, New York.

[7] Cover, T.M. and Thomas, J. 1991, '*Elements of Information Theory*', Wiely & sons, New York.

[8] Quinlan, J. R. 1993, *C4.5: programs for machine learning (Vol.1)*. Morgan kaufmann.

[9] Burrows, W. R., Benjamin, M., Beauchamp, S., Lord, E. R., McCollor, D., & Thomson, B. 1995, 'CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada' *Journal of applied meteorology*, 34(8), 1848-1862

[10] Horton, P., & Nakai, K. 1996, 'A probabilistic classification system for predicting the cellular localization sites of proteins'. *In Ismb, Vol. 4, pp. 109-115*.

[11] Han, J. and Kamber M., 2000, '*Data Mining: Concepts and Techniques*', Morgan Kaufmann, pp.164-172.

[12] Taneja, I .J. 2001, '*Generalized Information Measures and Their Applications*', on-line book www.mtm.ufsc.br/~taneja/book/book.html

[13] Chakrabarti A, Yaoyun S., Anthony W., and Andrew Y., 2001, 'Informational complexity and the direct sum problem for simultaneous message complexity' Foundations of Computer Science, 2001. *Proceedings. 42nd IEEE Symposium on*, pp. 270–278. IEEE, 2001.

[14] Yossef Ziv Bar, Jayram TS, Kumar R, and Sivakumar D, 2002, An information statistics approach to data stream and communication complexity in Foundations of Computer Science, 2002, *The 43rd Annual IEEE Symposium on*, pages 209–218. IEEE, 2002

[15] Burgin, M. *Super-Recursive Algorithms*; Springer: New York, NY, USA, 2005.

[16] Eberbach, E., (2005), 'Calculus of bounded rational agents: Flexible optimization as search under bounded resources in interactive systems'. *Fundam. Inform.* , 68, 47–102.

[17] Eberbach, E.(2008) Approximate reasoning in the algebra of bounded rational agents. *Int. J. Approx. Reason*. 2008, 49, 316–330, doi:10.1016/j.ijar.2006.09.014.

[18] Eberbach, E(2007). The Calculus process algebra for problem solving: A paradigmatic shift in handling hard computational problems. *Theory. Computer. Sci.* 2007, 383, 200–243, doi:10.1016/j.tcs.2007.04.012.

[19] George M. Marakas, (2004) '*Modern Data Warehousing, Mining, and Visualization: Core Concept'*, Prentice Hall/Pearson, chapter 3.5.

[20] Maszczyk, T. and Duch, W., 2008, June. Comparison of Shannon, Renyi and Tsallis entropy used in decision trees. *In International Conference on Artificial Intelligence and Soft Computing* (pp. 643-651). Springer, Berlin, Heidelberg.

[21] Ömer AKGÖBEK (2011), 'A new algorithm for knowledge discovery from data sets using cross-entropy measurement', *Scientific Research and Essays Vol*. 6(20), pp. 4301-4311.

[22] Papalexiou SM, Koutsoyiannis D(2011) 'Entropy based derivation of probability distributions' A case study to daily rainfall. *Adv Water Resources* (2011), doi:10.1016/j.advwatres.2011.11.007

[23] Dr. H. C. Taneja, Dr. Ashok Chitkara and Mr. Vikas Kumar' Classification of Census Using Information Theoretic Measure Based ID3 Algorithm' '*International Journal of Mathematical Analysis*, Vol. 6, 2012, no. 51, 2511 – 2518'

[24] Dr. H. C. Taneja, Dr. Ashok Chitkara and Mr. Vikas Kumar 'Census-2011 Using Modified ID3 Algorithm' '*National Conference on Mathematical Modeling, Simulation and Optimization'* at Chitkara University, Punjab.

[25] T.M. Cover and J.A. Thomas (2012), *Elements of Information Theory*, Wiley, 2012.

[25] D.S.Hooda and B.K.Hooda (2006), Dimension Reduction in multivariate analysis using maximum entropy criterion, *Journal of Statistics & Management systems* Vol.9 (2006),No 1,pp.175-103

[27] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah (2018), Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis, J. Phys.: *Conf. Ser., 983(1)*, p. 012063.

[28] L. Yi-bin, W. Ying-ying, and R. Xue-wen (2017), Improvement of ID3 algorithm based on simplified information entropy and coordination degree, Chinese *Automation Congress (CAC)*, pp. 1526–1530.

[29] M. A. Muslim, A. Nurzahputra, and B. Prasetiyo (2018) , Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction, *IEEE International Conference on ICT (ICOIACT)*, pp. 141–145.

[30] Z. Wang, Y. Liu, and L. Liu(2017), A new way to choose splitting attribute in ID3 algorithm, in 2017 IEEE 2nd Information Technology, Networking, *Electronic and Automation Control Conference (ITNEC)*, pp. 659–663

[31] Cherfi, A., Nouira, K., and Ferchichi, A. (2018). Very Fast C4.5 Decision Tree Algorithm. *Journal of Applied Artificial Intelligence*, 32(2), pp. 119-139

[32] A. Desai, S. Chaudhary (2017), Distributed decision tree v.2.0, in: *IEEE International Conference on Big Data (Big Data)*, 2017, pp. 929–934

[33] M. Li (2017), Application of CART decision tree combined with PCA algorithm in intrusion detection, *8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp. 38–41.

[34] Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang (2019), Prediction performance of improved decision tree-based algorithms ; *2nd International Conference on Sustainable Materials Processing and Manufacturing*, Procedia Manufacturing 35 698–703.

[35] Parmil Kumar & D. S. Hooda (2014) Information Theoretic Approach for Analyzing Independence of Attributes in Contingency Tables, *Journal of Statistics and Management Systems* , 17:2, 143-153, DOI: 10.1080/09720510.2013.857913

[36] D.S.Hooda and B.K.Hooda (2001), On Measurement of Stochastic Dependence in Multivariate Data, *Indian Journal of Pure and Applied Maths*, 32 (6): 801-815,June 2001.